The New Biology and Terascale Computing

November 13, 2001

William J. Camp, PhD

Director

Computers, Computation, and Mathematics

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy under Contract DE-AC04-94AL85000.

Sandia National Laboratories

A Perspective

Recently NAS convened a Panel on Partnerships in Computing & Biology .

It concluded that Biotech will soon be the dominant technology-- in terms of mindshare and R&D funding

It also concluded that Computing and Information Science, along with Nano and Micro-technologies, will open the door to the new biotechnology

Bill Camp
Bill@sandia.gov

Sandia National Laboratories

A Perspective

The NAS Panel on Partnerships in Computing & Biology :

also concluded that most professional biologists and MDs are not educationally prepared to deal with this new wave in their sciences.

Eventually biology and medical curricula (and the students they attract) will accommodate this new approach.

In the interim there is an enormous opportunity for physical scientists, CIS professionals and mathematicians to help jumpstart the field

Bill Camp
Bill@sandia.gov

Sandia National Laboratories

A Perspective

Research in the biological and computational sciences is increasingly complementary.

Informatics is a central component of many biotech businesses: Celera, Incyte,[include others]

Computational biology is becoming a recognized sub-domain of computer science

Bill Camp
Bill@sandia.gov

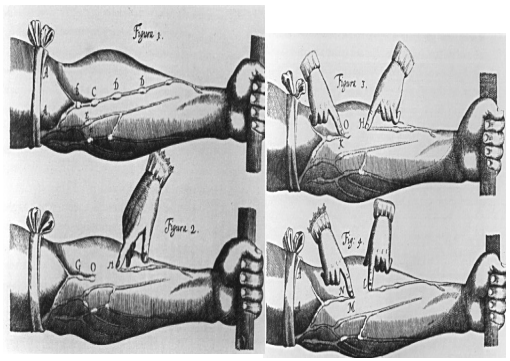
Sandia National Laboratories

Trends in Biology

Historically: an exploratory, descriptive, metaphorical, & empirical science

Detailed cataloguing and categorization of animal kingdom (Aristotle--> Linnaeus)

William Harvey's discovery of the circulation of blood.



Bill Camp
Bill@sandia.gov

Sandia National Laboratories

Trends in Biology (cont'd)

In the past 100 years: more discovery- and mechanism-oriented (but continuing to be empirical)

Mendellian Genetics

Models inherited phenotypic traits by indivisible units

Offspring inherit a copy of a trait's unit from each parent, and one of these two traits is expressed.

Darwinian Model of Evolution

Natural selection: process by which competition amongst individuals preserves "favorable" traits

-- leads to differentiation amongst species

Bill Camp
Bill@sandia.gov

Sandia National Laboratories

Current Trends in Biology

Biology is becoming a mechanistic and informational science

Human Genome Project:

fast sequencers -> explosion of sequence data

Functional Genomics
Molecular chemistry of cellular metabolism

Cell signalling pathways

Even more data!

Bill Camp
Bill@sandia.gov

Sandia National Laboratories

Current Trends in Biology

Biology and Medicine have tended to (fund and) carry out Hypothesis-driven research.

The Physical Sciences, CS and Math (the “hard” sciences) have advanced through Discovery-based Science.

An interesting question:

How would the War on Cancer advanced had DARPA rather than NIH funded it?

A hint: Biology funding is increasingly moving toward Discovery Science.

Bill Camp
Bill@sandia.gov

Sandia National Laboratories

Trends in Computer Science

Historically: a procedure oriented discipline based on Algorithms

Define a procedure to solve a problem

A problem is defined by algorithmic inputs and outputs

Bill Camp
Bill@sandia.gov

Sandia National Laboratories

Trends in Computer Science (cont'd)

In the past 20 years: more exploratory and empirical--

development of domain-specific heuristics that are empirically validated

Note: many interesting problems are intractable

Classical intractability and approximation results:

can guide our study of problems, but

do not provide practical solutions!

Bill Camp
Bill@sandia.gov

Sandia National Laboratories

Current Trends in CS

Commodity parallel computing (1 Gflop -> multi Tflops)

We now have the computational capacity to work with large data sets and attack difficult problems

Computational science

Applies mathematics and computer science to physical models to simulate reality

In biology: quantum chemistry; molecular dynamics; continuum models and heuristics for informatics

Bill Camp
Bill@sandia.gov

Sandia National Laboratories

In Sum:

Computer Science has historically been a procedure-oriented discipline

Biology has historically been descriptive, encyclopedic, metaphorical and historical.

Biology is becoming informational and more mechanistic

These trends in Biology are driven by high-throughput experimentation and a new discovery-science emphasis

In the past 20 years, Computer Science has become more exploratory and empirical (many interesting problems are intractable).

These trends in Computer Science have been accompanied by huge advances in computing capabilities--*commodity parallel computing* (1 Gflop to multi-Tflops)

Bill Camp
Bill@sandia.gov

Sandia National Laboratories

So where are we?

We are at a ripe time for collaboration of biology and computing

Biology has a lot of data to work with (bio-informatics)

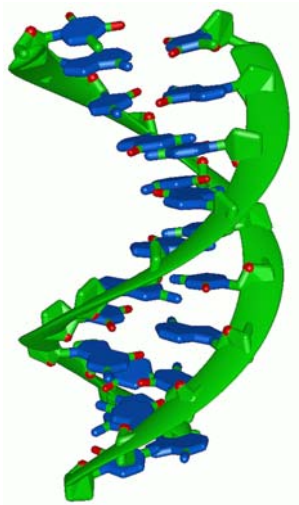
For the first time, adequate computational resources are available to tackle realistic biological problems

Simulation and Informatics for Biology and Computing will replace Physics Chemistry as the Driving Application of High Performance Computing

Bill Camp
Bill@sandia.gov

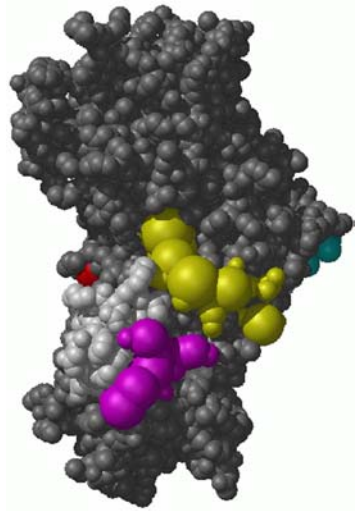
Sandia National Laboratories

Computational Opportunities in Cellular Biology



The Central Dogma of Biology

DNA⇒RNA⇒Protein



drives a need to translate the burgeoning amount of genomic data to action in individual cells.

Because of the complexity of the systems involved (metabolic, transport, signaling, …) this translation process can only occur with the assistance of computational methods.

<u>Some of the Necessary Computational Capabilities</u>

Molecular Biophysics

Complex System Modeling

Information Analysis

Bill Camp
Bill@sandia.gov

Sandia National Laboratories

What's Happening with High End Computing?

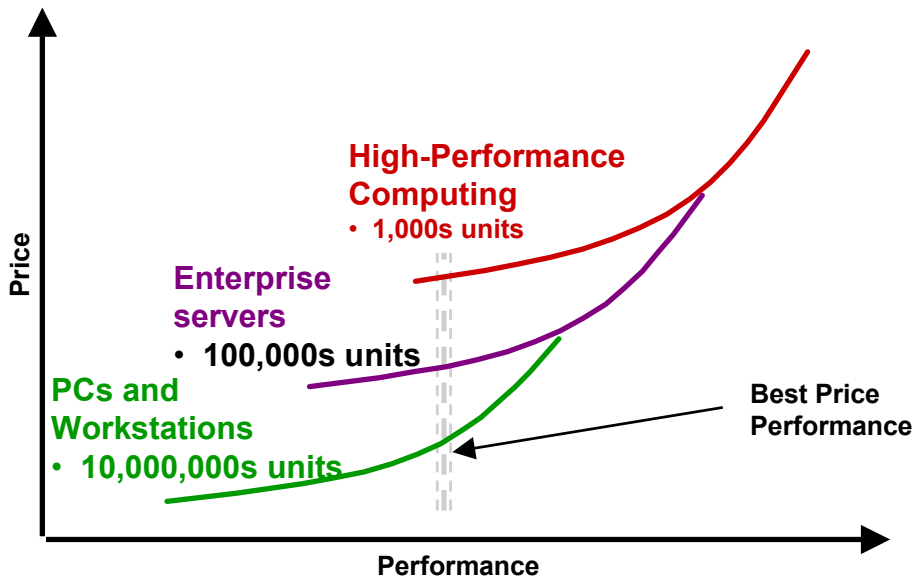
- High-volume building blocks available
- Commodity trends reduce cost
- Assembling large clusters is easier than ever
- Incremental growth possible
- HPC market is small and shrinking (relatively)
- Performance of high-volume systems increased dramatically
- Web driving market for scalable clusters
- Hot market for high-performance interconnects

(Commodity, Commodity, Commodity)

Bill Camp
Bill@sandia.gov

Sandia National Laboratories

Commodity Value Propositions



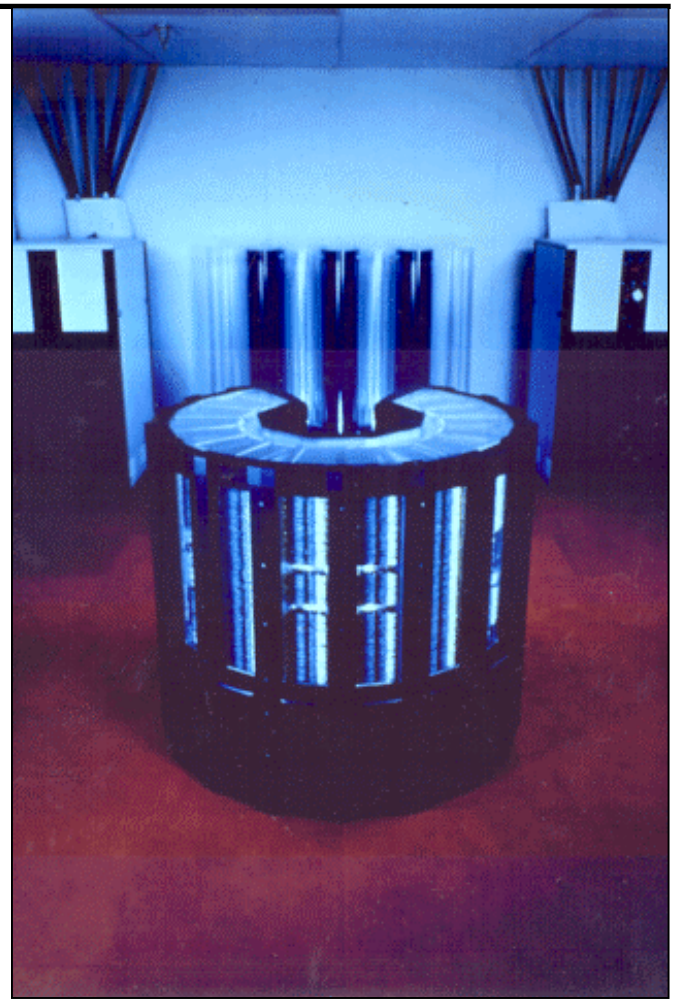
- Low cost to drive down the cost of simulation
- Flexible and adaptable to changing needs
- Manage and operate as a single distributed system

High Volume Technologies Give Favorable Price-Performance

Bill Camp
Bill@sandia.gov

Sandia National Laboratories

Cray-1



Bill Camp
Bill@sandia.gov

Sandia National Laboratories

nCUBE-2 Massively Parallel Processor (1024)

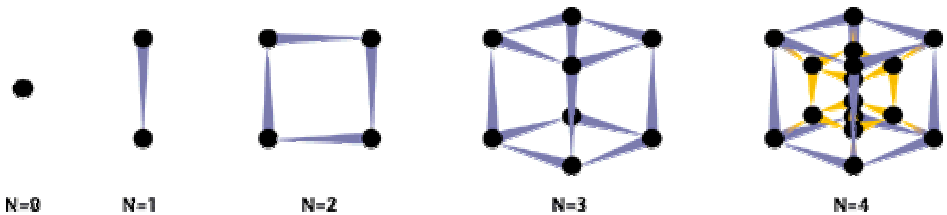
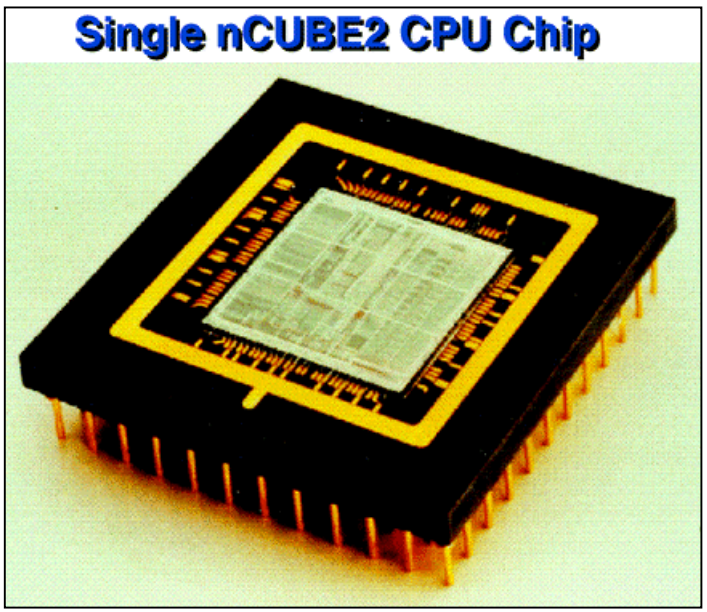
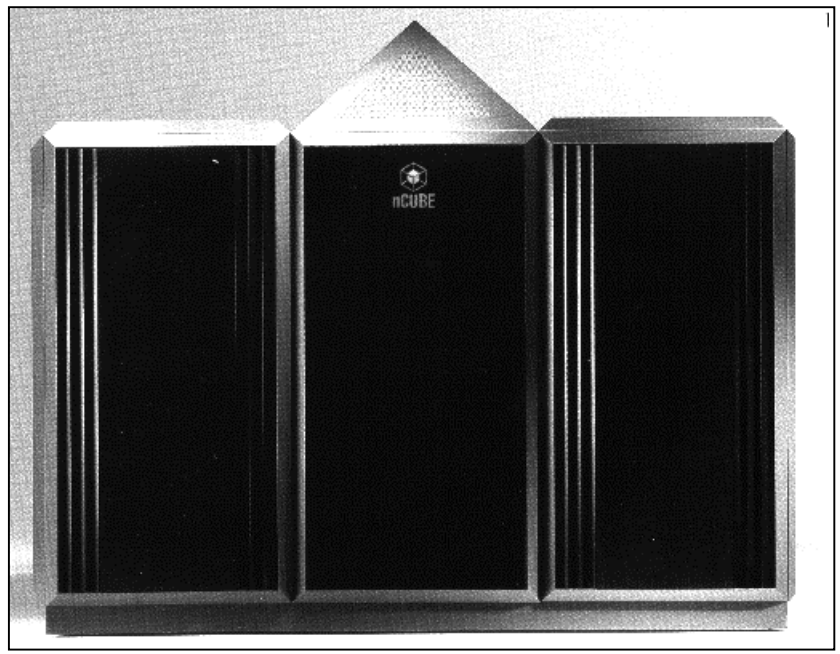


Figure 1: Two hypercubes of the same dimension, joined together, form a hypercube of the next dimension. N is the dimension of the hypercube.





Bill Camp
Bill@sandia.gov

Sandia National Laboratories

Intel Paragon

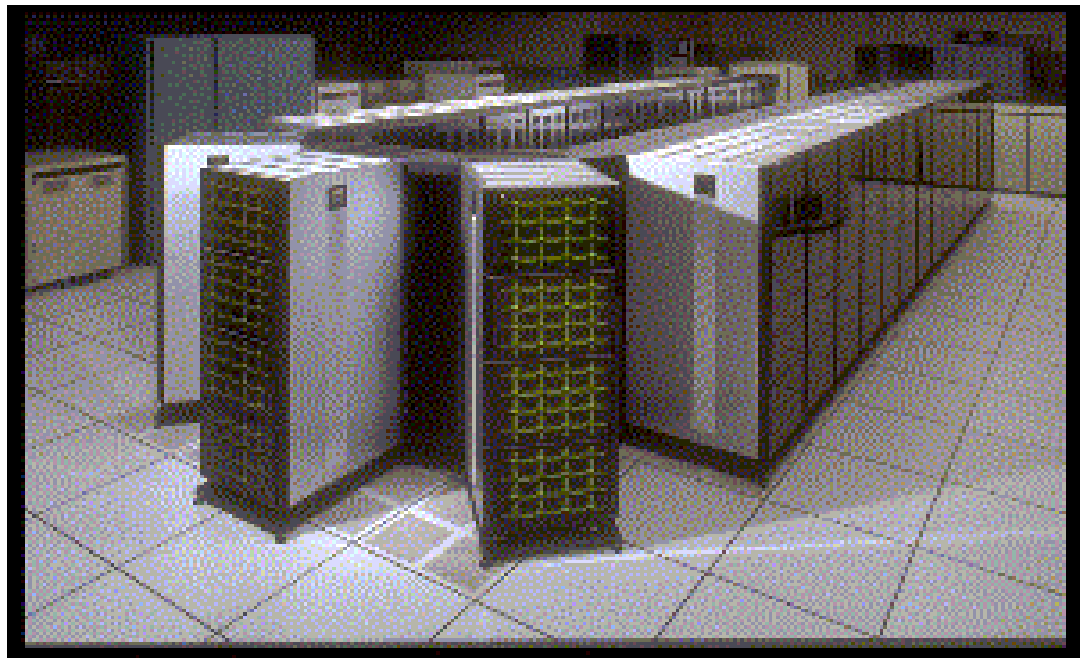
1,890 compute nodes

3,680 i860 processors

143/184 GFLOPS

175 MB/sec network

SUNMOS lightweight kernel



Bill Camp
Bill@sandia.gov

High Performance Computing at Sandia: Hardware & System Software

ASCI Red: The World's First Teraflop Supercomputer

9,472 Pentium II processors

2.38/3.21 TFLOPS

400 MB/sec network

Puma/Cougar lightweight kernel OS

Cplant™: The World's Largest Commodity Cluster

~2,500 Compaq Alpha Processors

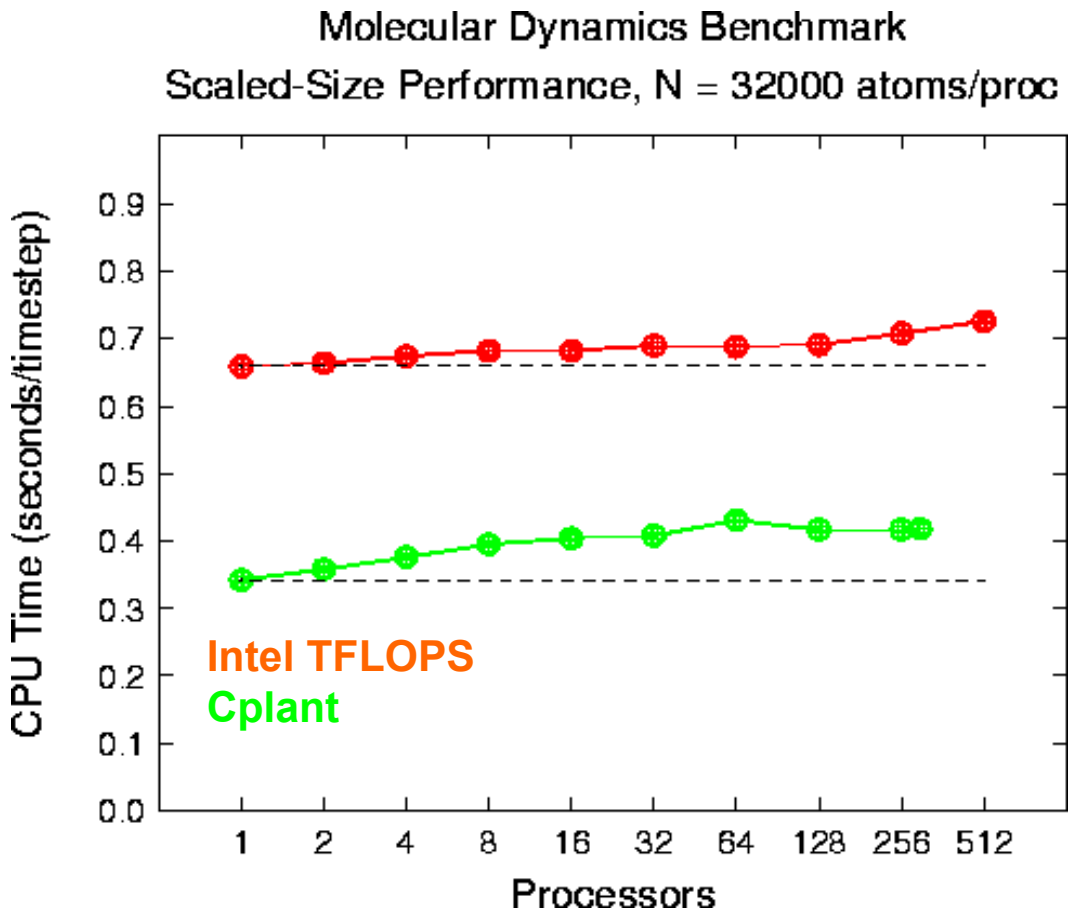
~2.7 Teraflops

Myrinet Interconnect Switches

Bill Camp
Bill@sandia.gov

Sandia National Laboratories

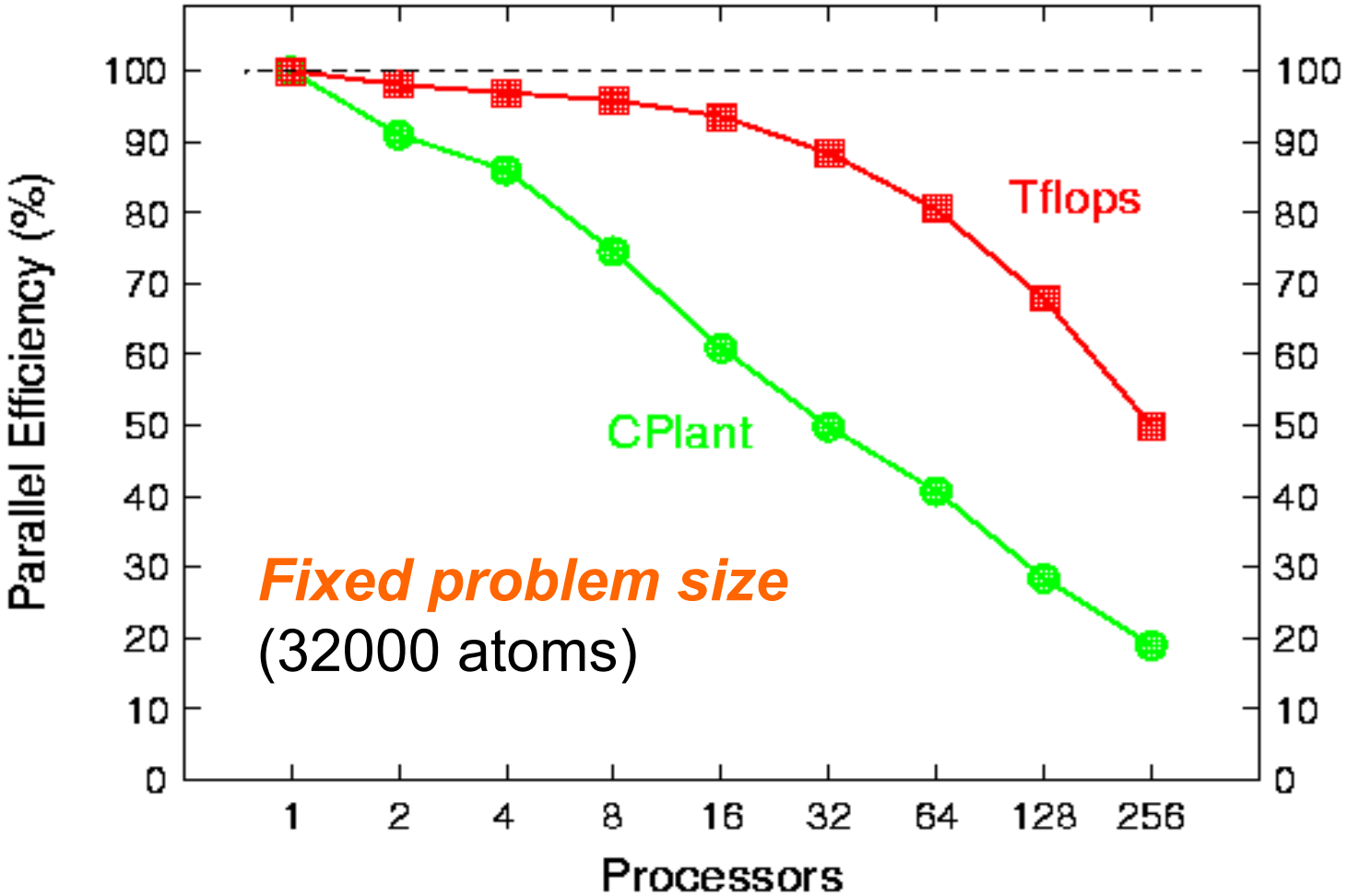
Cplant Performance Relative to ASCI Red



Bill Camp
Bill@sandia.gov

Sandia National Laboratories

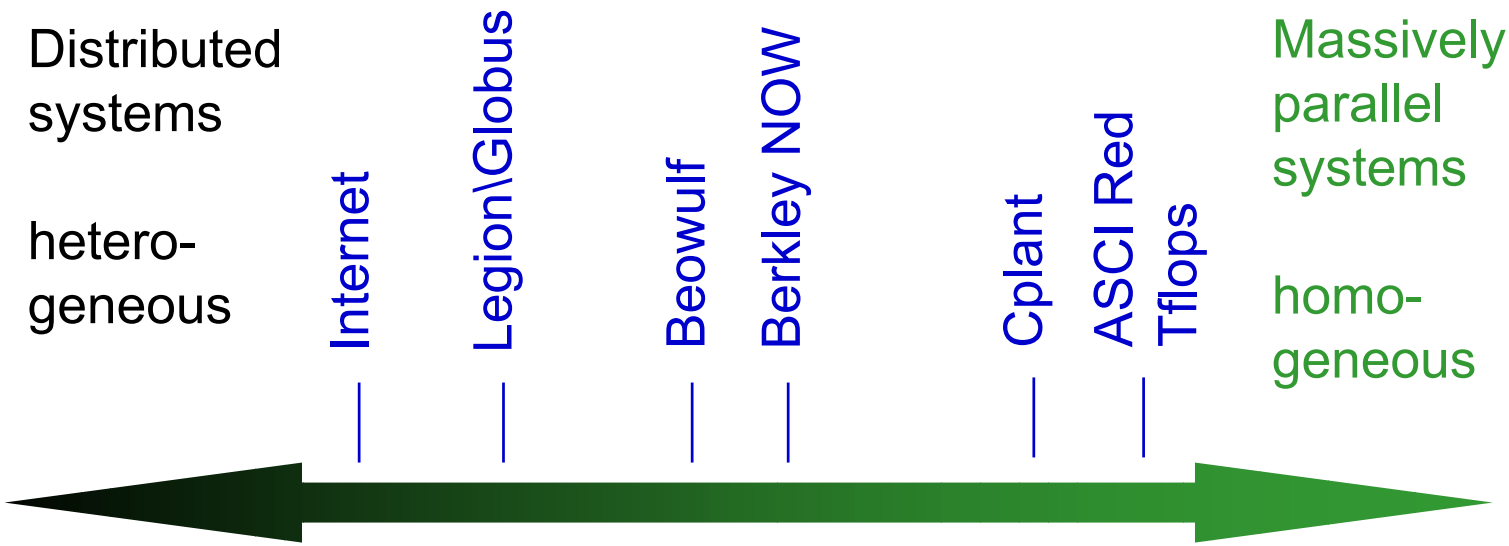
Molecular Dynamics Benchmark (LJ Liquid)



Bill Camp
Bill@sandia.gov

Sandia National Laboratories

Distributed & Parallel Systems



Gather (unused) resources
Steal cycles
System SW manages resources
System SW adds value
10% - 20% overhead is OK
Resources drive applications
Time to completion is not critical
Time-shared

Bounded set of resources
Apps grow to consume all cycles
Application manages resources
System SW gets in the way
5% overhead is maximum
Apps drive purchase of equipment
Real-time constraints
Space-shared

Bill Camp
Bill@sandia.gov

Sandia National Laboratories

Communication-Computation Balance for Past and Present Massively Parallel Supercomputers

Machine	*Balance Factor (bytes/s/flops)*
Intel Paragon	1.8
Ncube 2	1.0
Cray T3E	0.8
ASCI Red	0.6
Cplant™	0.1

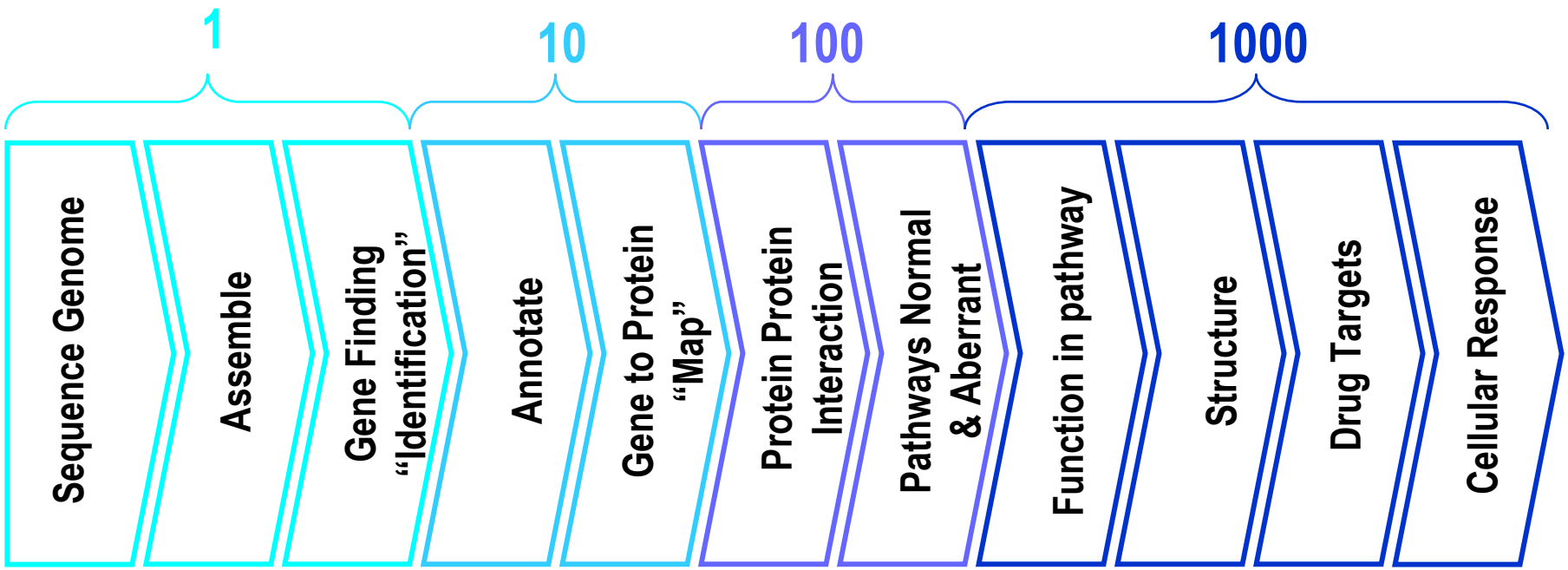
But Does Balance Matter for Computational Biology ?

Bill Camp
Bill@sandia.gov

Sandia National Laboratories

Computational Pharmaceutical Development

A Guess at the Computing Power Needed ***(in TeraOps)***



"Embarassingly Parallel"

"Massively Parallel"

Bill Camp
Bill@sandia.gov

Sandia National Laboratories

Assemble Genome

Can only sequence 500-600 base pairs at a time: need to combine them into full genome

- First, find overlaps between strings (string matching with errors, dynamic programming and other advanced algorithms).
- Then merge strings using overlap information (have to handle repeats, gaps and errors: various sophisticated algorithmic insights

Bill Camp
Bill@sandia.gov

Sandia National Laboratories

Gene Finding, Annotation, Gene to Protein Map:
Sequence Comparisons

Given new sequence, what's it similar to?

Find putative genes in new sequence

Gain insight into function of new sequence

Inexact string matching

Base pairs or amino acids

Large strings & libraries so fast algorithms essential

Need to quantify quality of match

BLAST, FASTA & lots of new algorithm development

Bill Camp
Bill@sandia.gov

Sandia National Laboratories

Gene Finding, Annotation, Gene to Protein Map:
Gene Finding

Given a genome, find the genes

Only small fraction of mammalian genome is genes

Some patterns are understood

e.g. start and stop of coding regions

But post-transcription editing is poorly understood

Much gets spliced out before becoming protein

String-based techniques have only modest success

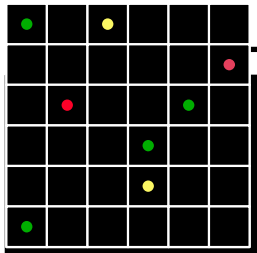
RNA secondary or tertiary structure might be key?

Bill Camp
Bill@sandia.gov

Sandia National Laboratories

VXInsight™ Analysis of Microarray Data

Given the enormous amount of data produced by today's and tomorrow's high-throughput experimental biology methods, defining and developing an adequate data visualization and analysis environment is essential.



Bill Camp
Bill@sandia.gov

Sandia National Laboratories

Functional Proteomics

Identify functional relationships between proteins through large-scale experimental assays

- Empirical Analysis
- Data mining and pattern recognition
- Methods for large-scale data sets
- Methods to account for noise/bias in the data

Bill Camp
Bill@sandia.gov

Sandia National Laboratories

Protein Structure

Motivation

- A protein's structure provides strong clues about its function
- Known protein structures are invaluable for drug design (through docking experiments)

The Challenge

- There are over 1 million proteins.
- Currently know the tertiary structures of 1000s of proteins
- Difficult to combine different sources of experimental data (e.g. distance constraints and homology)

Bill Camp
Bill@sandia.gov

Sandia National Laboratories

Limitations of Experimental Methods

Crystallography

- Crystallized protein structure (or complex) is used to determine the structure
- Very labor intensive
- Not feasible in all cases (flexible proteins, membrane-bound proteins)

NMR

- Proton resonances are measured to provide distance and local geometry constraints
- Fidelity of data limits application to at best moderate size proteins

Bill Camp
Bill@sandia.gov

Sandia National Laboratories

Computational Approaches

Computational Analysis

Lattice-based models

Intractability results and approximation algorithms

Methods from “exact” statistical mechanics (e.g. ab initio MD)

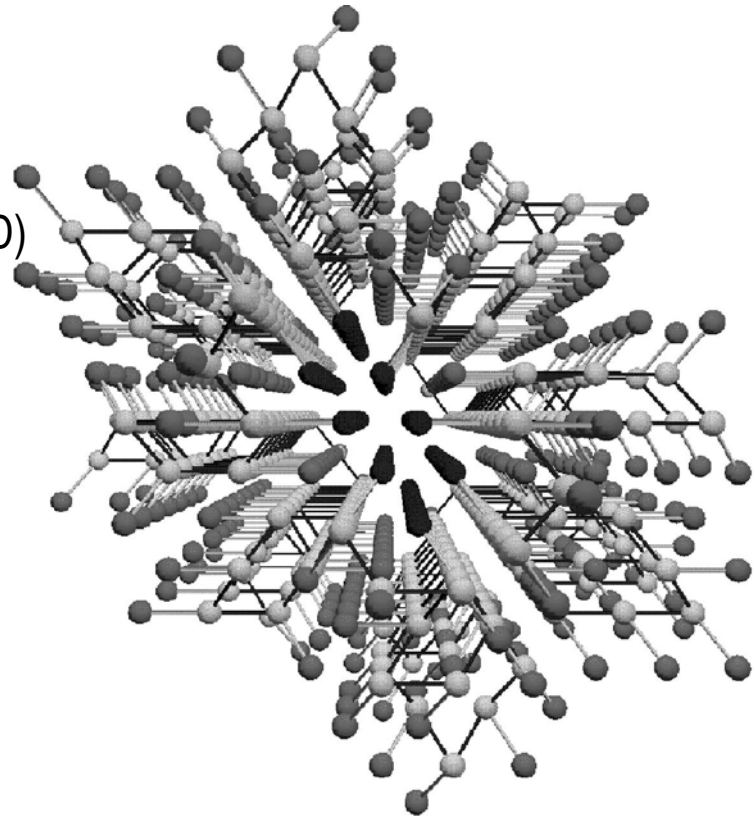
Global optimization-Search through a parameterized space of conformations & empirical energy potentials capture potential energy and approximate entropic effects

Homology & Threading

Exploit structure of known proteins

Large-scale sequence or conformational similarity assessment

Substructure alignment



Bill Camp
Bill@sandia.gov

Sandia National Laboratories

Constraint-Based Structure Determination

- Most membrane proteins cannot be crystallized and are too big for NMR
- Solve computationally: experimental distance constraints is the most promising technique for proteins which have little other information known about them.
  - Determination of interpeptide distance using mass spectroscopy and labeling NMR is becoming a high throughput technique:
  - Energy Evaluation + refinement with molecular physics methods.

Focusing first on satisfying the distance constraints yields an optimization problem that is amenable to mathematics and computer science techniques.

Given a reasonable number of constraints, proteins structure determination with this method will take about 0.1 - 10 TFLOP days.

Bill Camp
Bill@sandia.gov

Sandia National Laboratories

Structure Optimization

- Starting with crude coordinates, MD methods can get fairly precise coordinates and solvations.
- 1000 critical pathway steps with 10 alternate protein configurations times .003 of a TFLOP-month per structure is about 2.5 TFLOP-years.
- Increased accuracy (requiring more computing power) can be obtained via
  - Next generation force-fields or
  - A hierachical approach employing refinement with more accurate methods (e.g. quantum mechanics)

Bill Camp
Bill@sandia.gov

Sandia National Laboratories

Computational Limits of MD

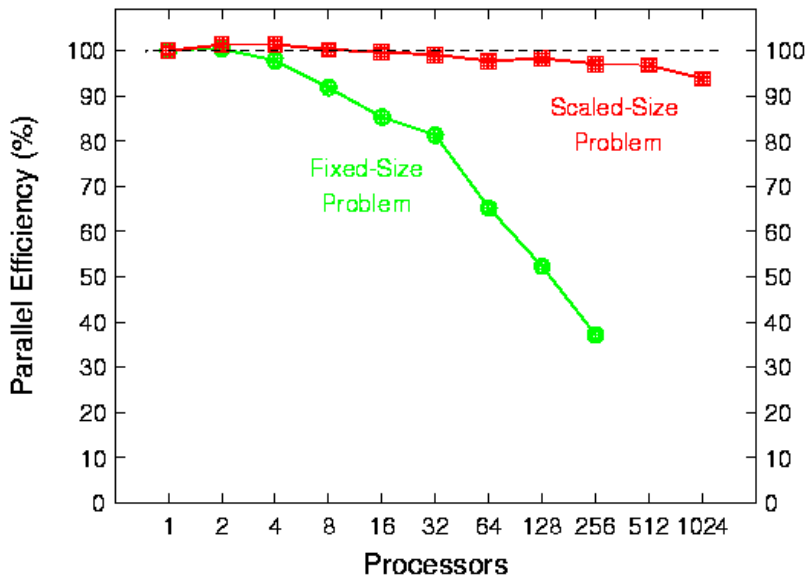
Fundamental limit: Time isn't parallel

- *length-scale = Angstroms*
- *timestep = femtoseconds*

e.g. Biomembrane simulation:

- **7134 atoms/proc (Intel Tflops)**
- **Long-range Coulombics via particle-mesh Ewald and parallel FFTs**

Supercomputer-class computation: tens of nanoseconds of 20-50,000 atoms



Bill Camp
Bill@sandia.gov

Sandia National Laboratories

Target Identification

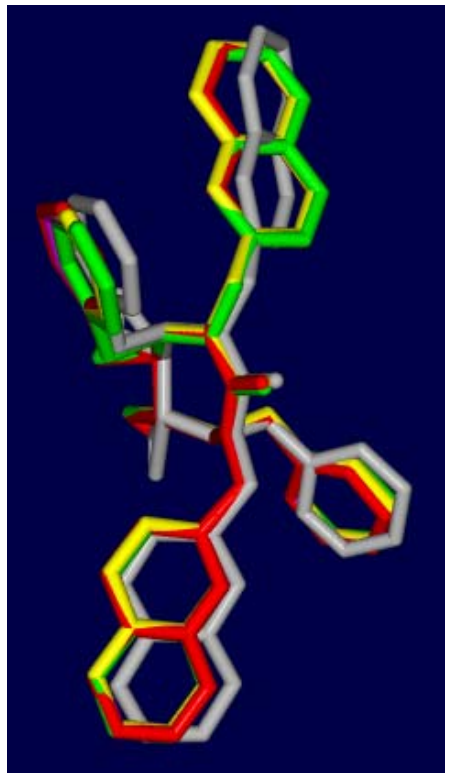
Computational screening of a million compounds for docking affinities with enzymes is also computational intense.

Goal: design or identify ligands that bind with a target macromolecule at a given site

Computational screening of one million compounds for docking affinities with enzymes is computationally intense:

For each drug target

- –1000 critical pathway steps (e.g. docking)
- –1 million small molecule compounds
- –10 alternate configurations for the enzyme
- –@ 0.01 of a TFLOP-month per enzyme
- –= 100 TFLOP-months or *8 TFLOP-years per drug target.*



New Chemical Descriptors
- Enable rapid searches of ligand databases with *known* accuracy.

Rigid docking
- Can quickly evaluate binding potential for a ligand
- Large-scale search for chemical databases

Flexible docking
- Global optimization to find ligand position and configuration
- Can use more detailed energy potentials

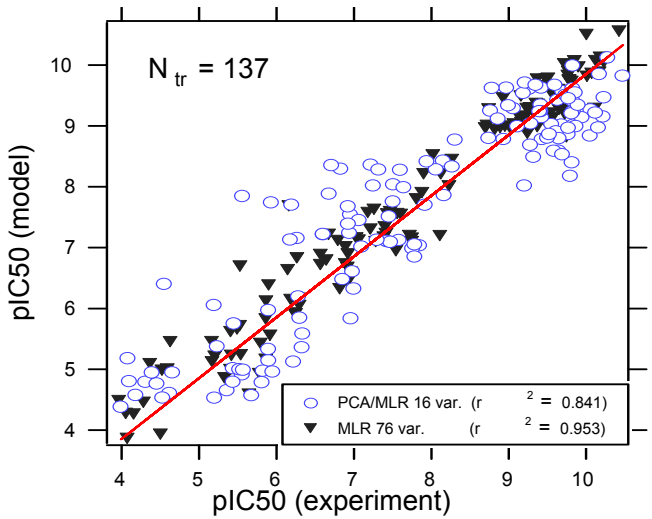
Bill Camp
Bill@sandia.gov

Sandia National Laboratories

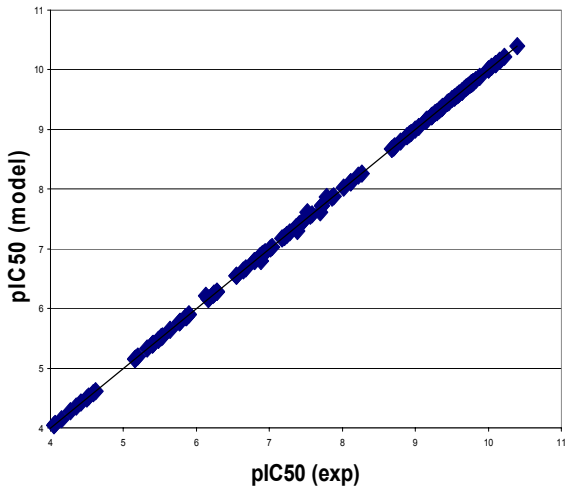
One Improvement: New Database Methods for Structure-Property Relationships

QSAR Equation with Signature

HIV-1 protease inhibitors binding affinities (pIC50)



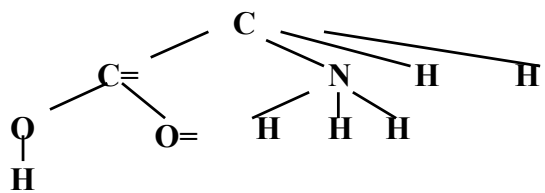
Typical QSAR (Molconn-Z descriptors)



Signature descriptors (extended connectivity index)

Glycine

$(H_3N\text{-}C_\beta H_2\text{-}COOHH)$



Bill Camp
Bill@sandia.gov

Sandia National Laboratories

Target Physics & Chemistry

Computational Molecular Biophysics

Molecular Simulation

- Molecular Dynamics (MD)
  - NVT, NVE
  - Grand Canonical MD
  - Reaction-ensemble MD
- Monte Carlo (MC)
  - Grand Canonical MC
  - Configurational Bias MC
  - Gibbs-ensemble MC
  - Transition state theory

Molecular Theory

- Classical Density Functional Theory

Electronic Structure Methods

- Local Density Approximation (LDA)
- Quantum Chemistry (HF etc.)

Mixed Methods

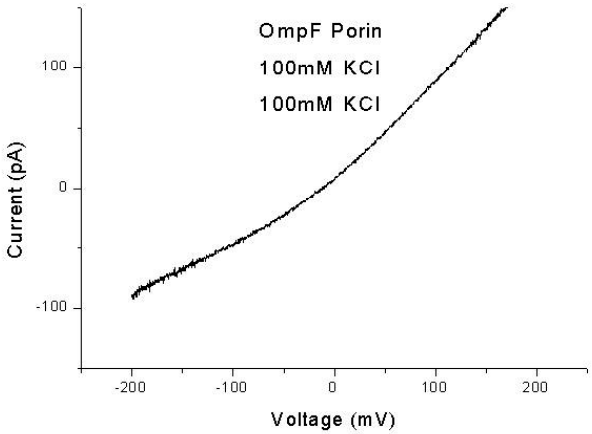
- Quantum-MD
- Quantum-CDFT
- Brownian Dynamics-CDFT

Bill Camp
Bill@sandia.gov

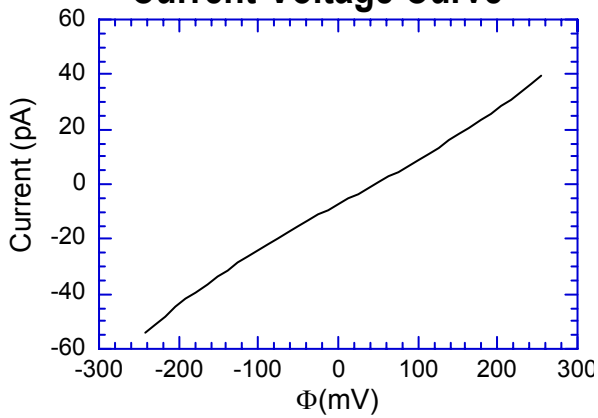
Sandia National Laboratories

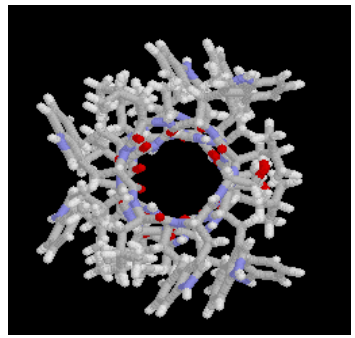
The Application of Laboratory Molecular Physics Capabilities to Biology Problems

Measured Ion Channel Current-Voltage Curve



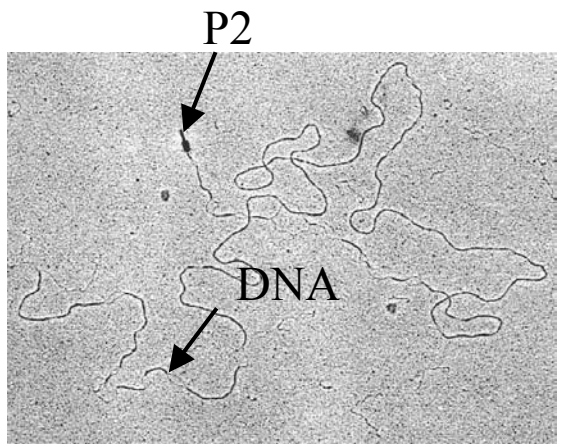
Calculated Ion Channel Current-Voltage Curve

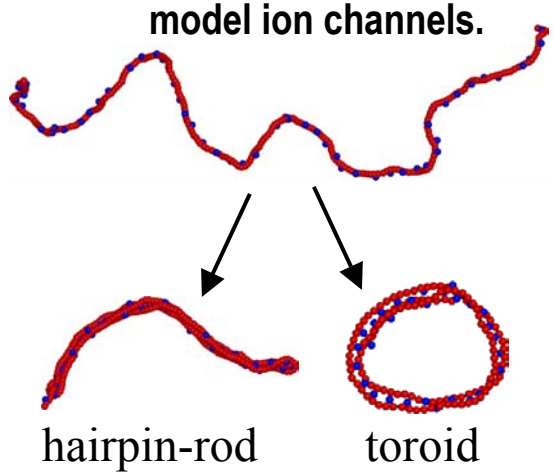




Development and application of molecular theory to model ion channels.

MP molecular dynamics simulations aimed at understanding how DNA (which orders of magnitude longer than any dimension of its host) fits into its host.

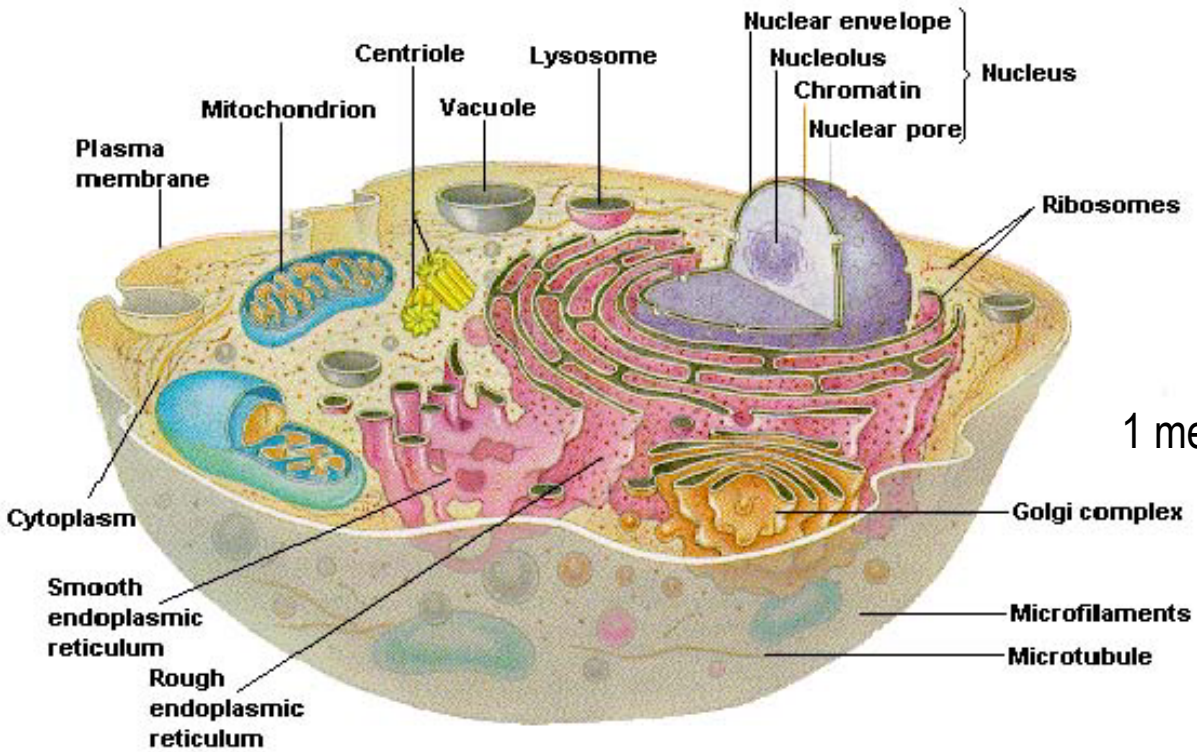




Bill Camp
Bill@sandia.gov

Sandia National Laboratories

What Data is Needed to Specify a Single Eukaryotic Cell?



Taken from *Human Biology* by Daniel Chiras

www.people.virginia.edu/~rjh9u/cell1.html

Organelles

4 Million Ribosomes

30,000 Proteasomes

Dozens of Mitochondria

Macromolecules

5 Billion Proteins

5,000 to 10,000 different species

1 meter of DNA with Several Billion bases

60 Million tRNAs

700,000 mRNAs

Chemical Pathways

Vast numbers

Tightly coupled

Is a Virtual Cell Possible?

Bill Camp
Bill@sandia.gov

Sandia National Laboratories

Virtual Cell Project

This is the hardest simulation challenge I have ever seen!

E. G.

~ 10**9 ATP molecules in a cell, They take about a second to turn over to ADP and back in a typical eukaryote.

You cannot model all the molecules. Nor can you ignore molecular detail:

These *are* molecular-scale machines.

Bill Camp
Bill@sandia.gov

Sandia National Laboratories

Virtual Cell Project

NCRR-funded center within the UConn Health Center, Center for Biomedical Imaging Technology

National Resource for Cell Analysis & Modeling (Virtual Cell)

http://www.nrcam.uchc.edu/

Sandia's Contributions

- efficient parallel implementation
- solving systems of stiff linear PDE's
- converting digitized images in 3d to meshable geometry

Bill Camp
Bill@sandia.gov

Sandia National Laboratories

Transport/Reaction Simulation of Fertilization Ca^{2+} Wave in Xenopus laevis eggs with MPSalsa

Ionic (Calcium) and Protein (Endoplasmic Reticulum IP_3R) Species

IP_3R Ion Channels Responding to Fertilization Event Initiates Calcium Wave Propagating through Egg

Concave Wave Front Modulated via Distributions of Protein Structures

Bistability of Calcium Concentrations pre- and post- Fertilization

Collaboration: VCELL Modeling Effort, University of Connecticut

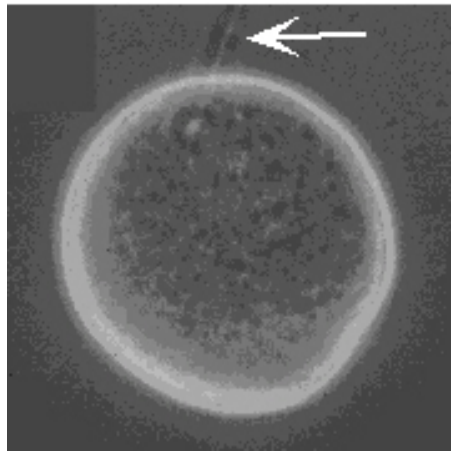


Image of Xenopus egg During fertilization event[1]

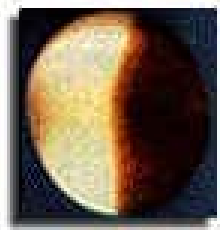


Image of Calcium wave following fertilization[2]

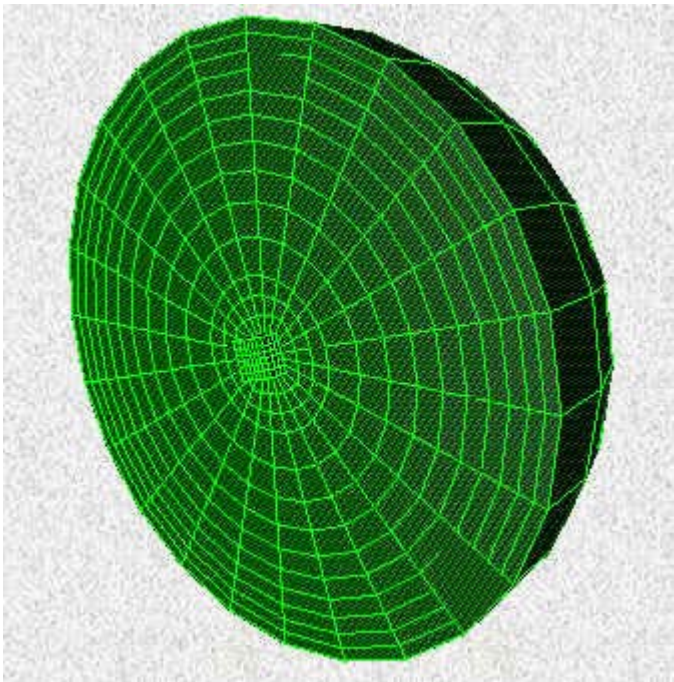
[1]*Image from: SPERM DISINTEGRINS, EGG INTEGRINS, AND OTHER CELL ADHESION MOLECULES OF MAMMALIAN GAMETE PLASMA MEMBRANE INTERACTIONS, Janice P. Evans, Frontiers in Bioscience 4, d114-131, January 15, 1999*

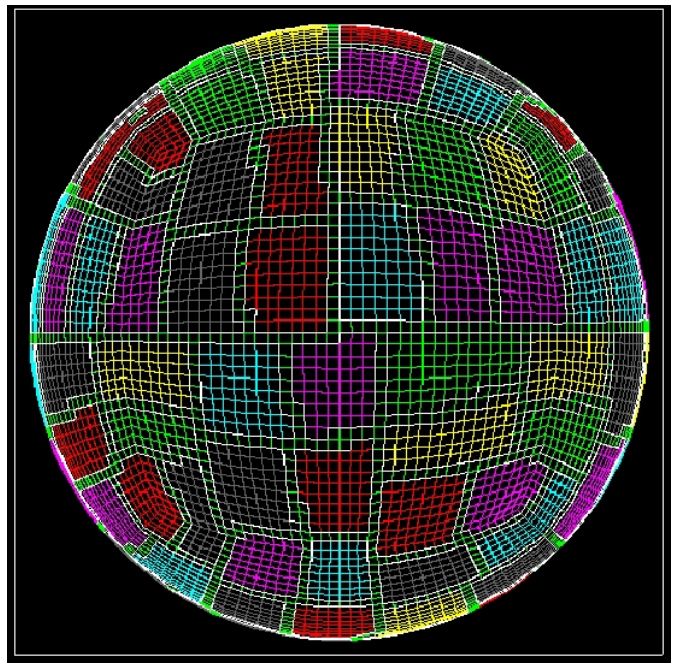
[2]*Image from:Amphibian Embryology Tutorial, http://worms.zoology.wisc.edu/frogs/fert/fert_oogen.html*

Bill Camp
Bill@sandia.gov

Sandia National Laboratories

Spatial Discretization (Meshing) of Spherical Egg Representation and Partition into Subdomains





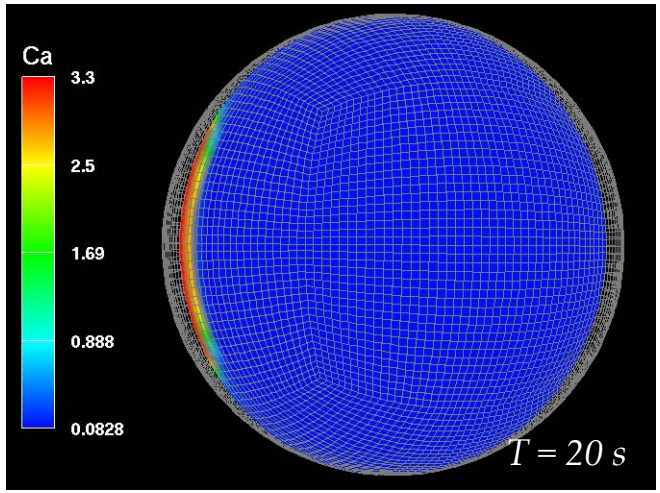
The egg is represented as a 500 μm radius sphere. Ilustrated above on the left is how the spherical region is meshed: Radially symmetric partitions toward the outer region and the surface, with a cube of rectangular elements at the center. This resulted in 371,200 elements and 376,185 of nodes. The mesh was partitioned into 512 subregions for solution on 512 processors – the partitioning is illustrated above on the right (surface view).

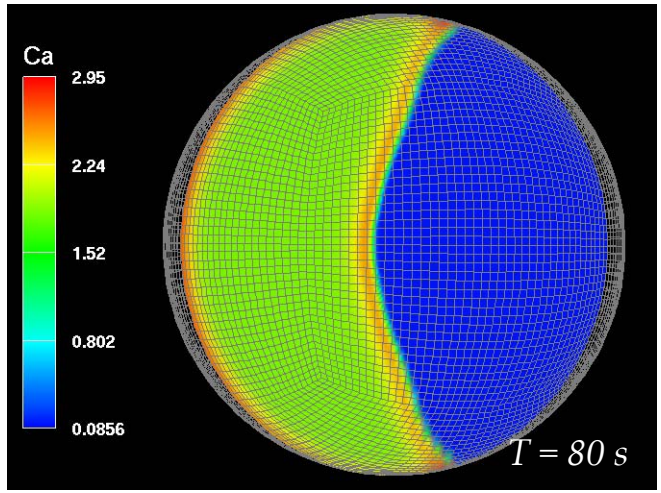
Bill Camp
Bill@sandia.gov

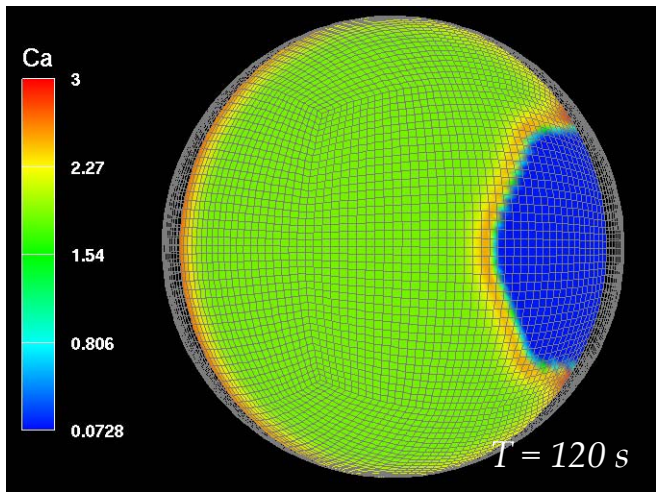
Sandia National Laboratories

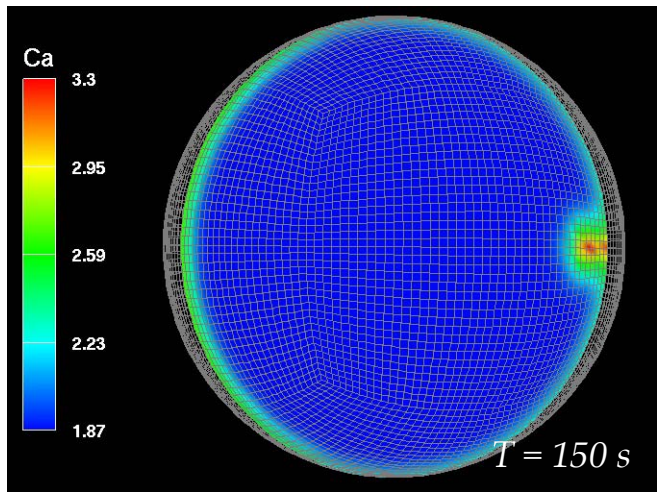
First fully 3-D simulation of the Ca^{2+} wave transport and reaction during fertilization of the Xenopus laevis egg

The egg is represented as a 500 mm radius sphere meshed with 371,200 elements and partitioned into 512 subregions for solution on 512 processors.





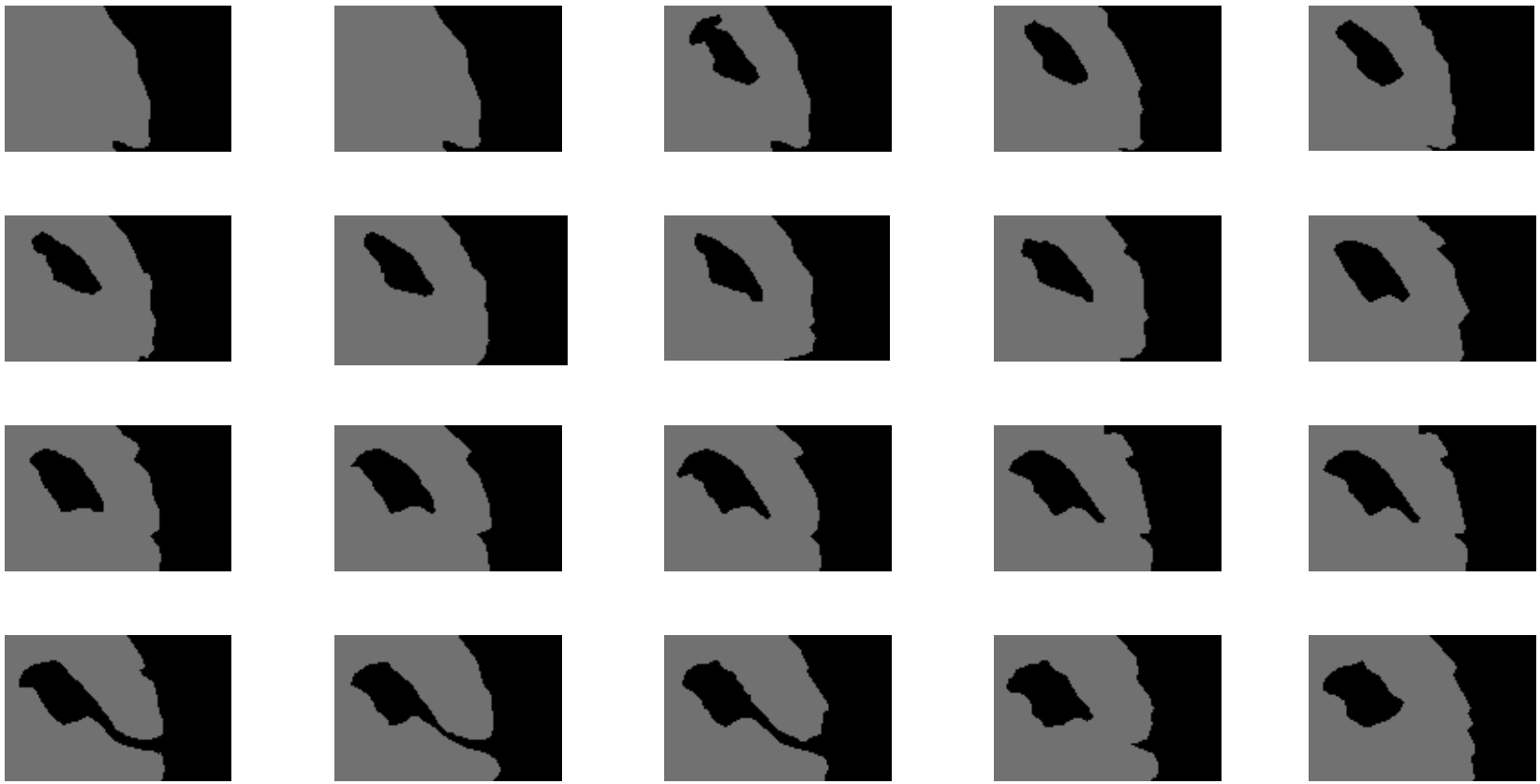




Bill Camp
Bill@sandia.gov

Sandia National Laboratories

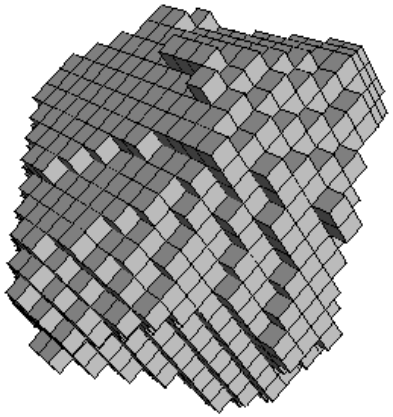
Mitochondria Cristae



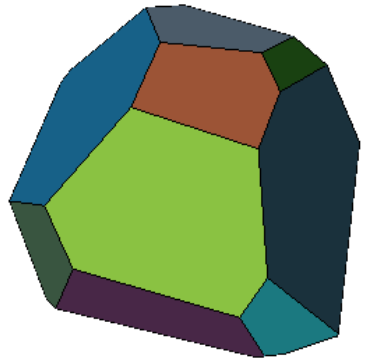
Bill Camp
Bill@sandia.gov

Sandia National Laboratories

Converting On-Lattice Grain Representation to Meshable Polyhedra



Lattice Algorithms for Grain Edge/Face Reduction (LAGER)

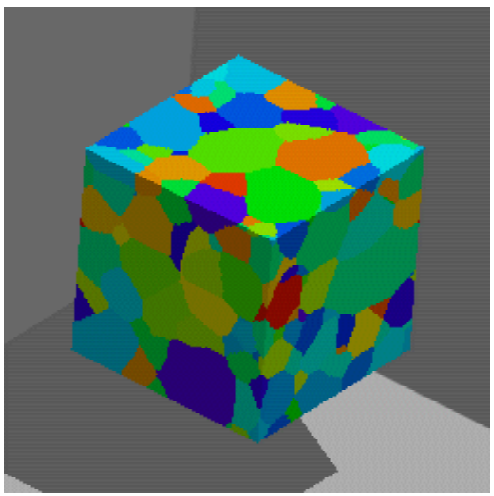


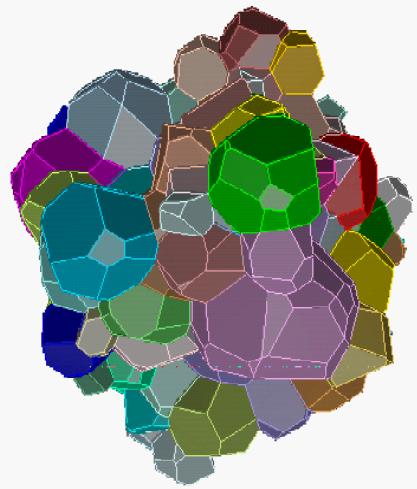
Convert from grain-growth methods or digitized experimental measurements into polyhedra for FEM.

Dramatically reduced memory requirements

More realistic/No lattice effects

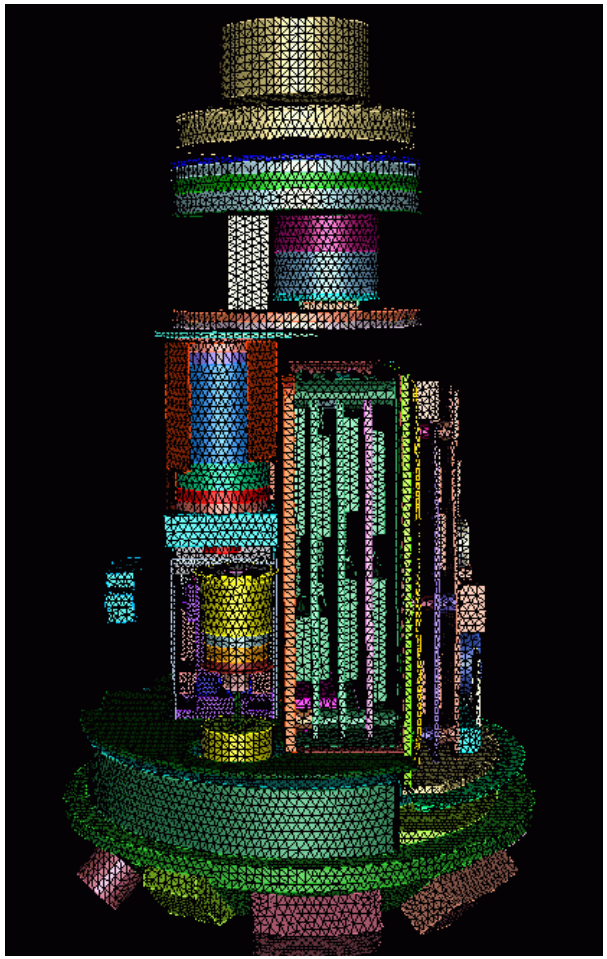
***O(n)* algorithm for narrow grain size distribution, *O(n ln n)* otherwise**

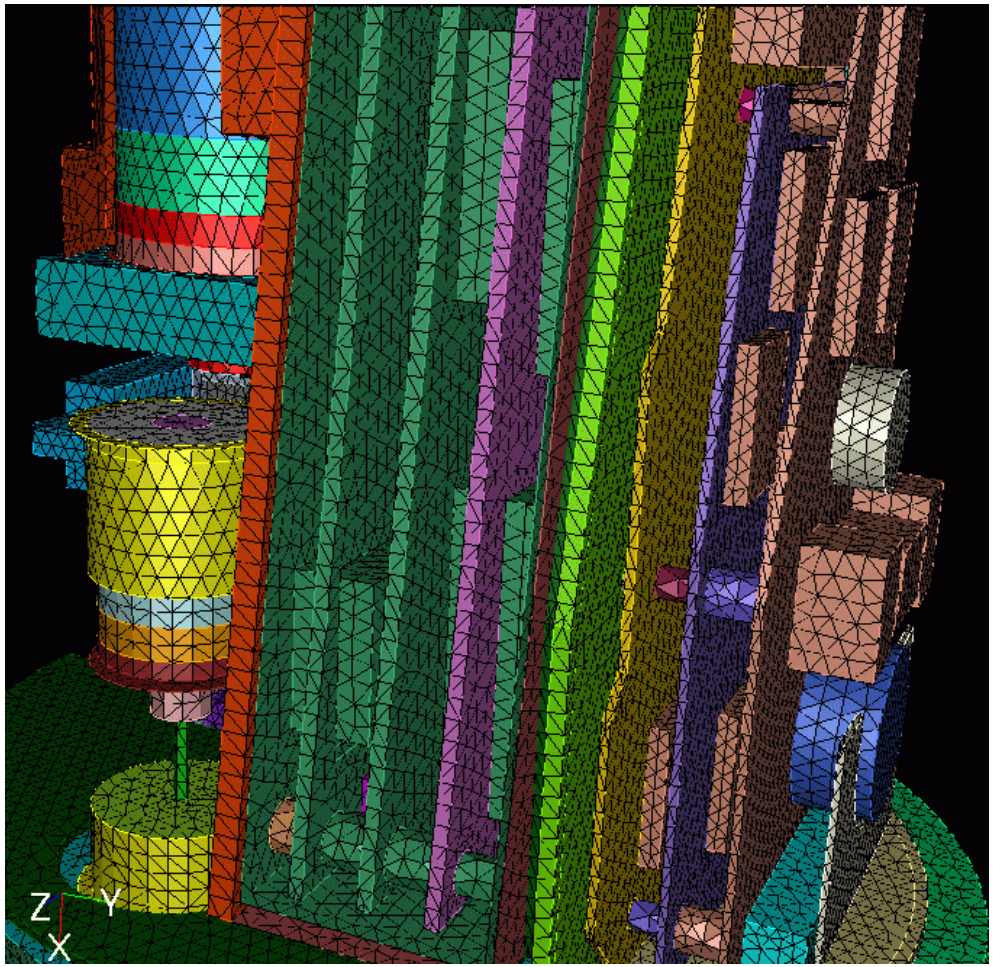




Bill Camp
Bill@sandia.gov

W76 AF&F mesh

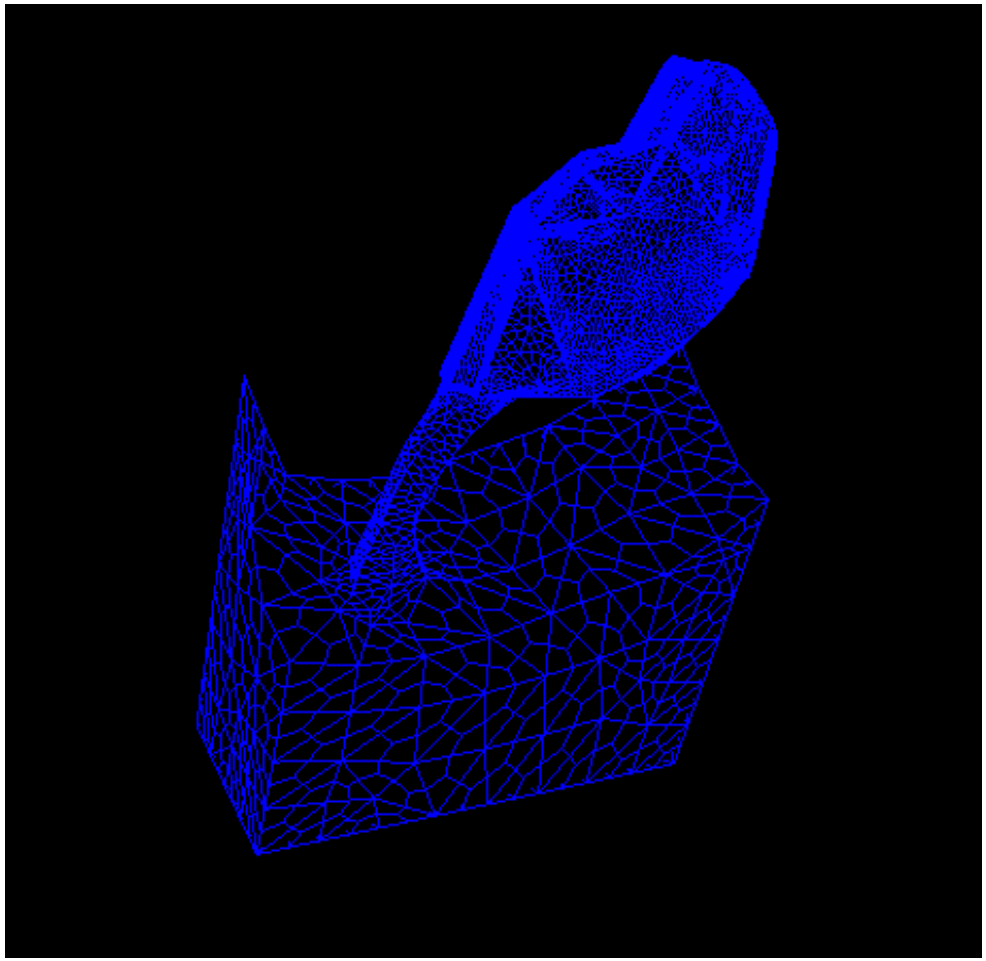




Bill Camp
Bill@sandia.gov

Sandia National Laboratories

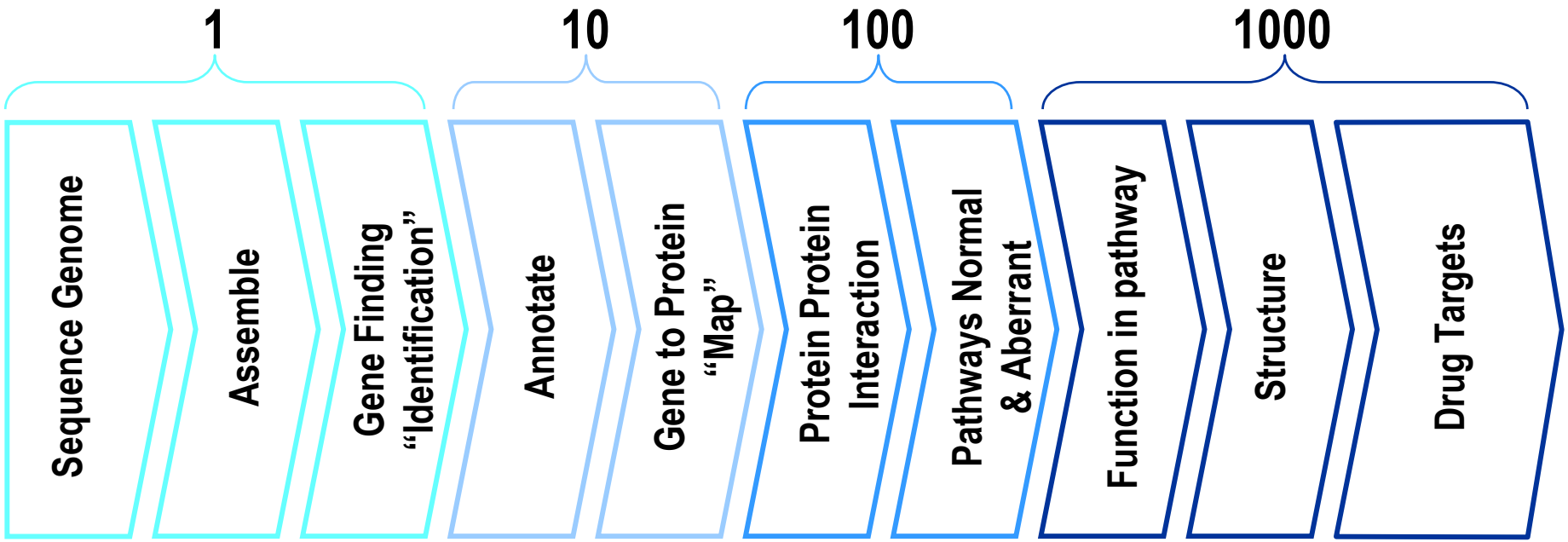
Meshing of mitochondrial cristae for 3-D ADP/ATP transport study within mitochondria



Bill Camp
Bill@sandia.gov

Sandia National Laboratories

Current Timeline?



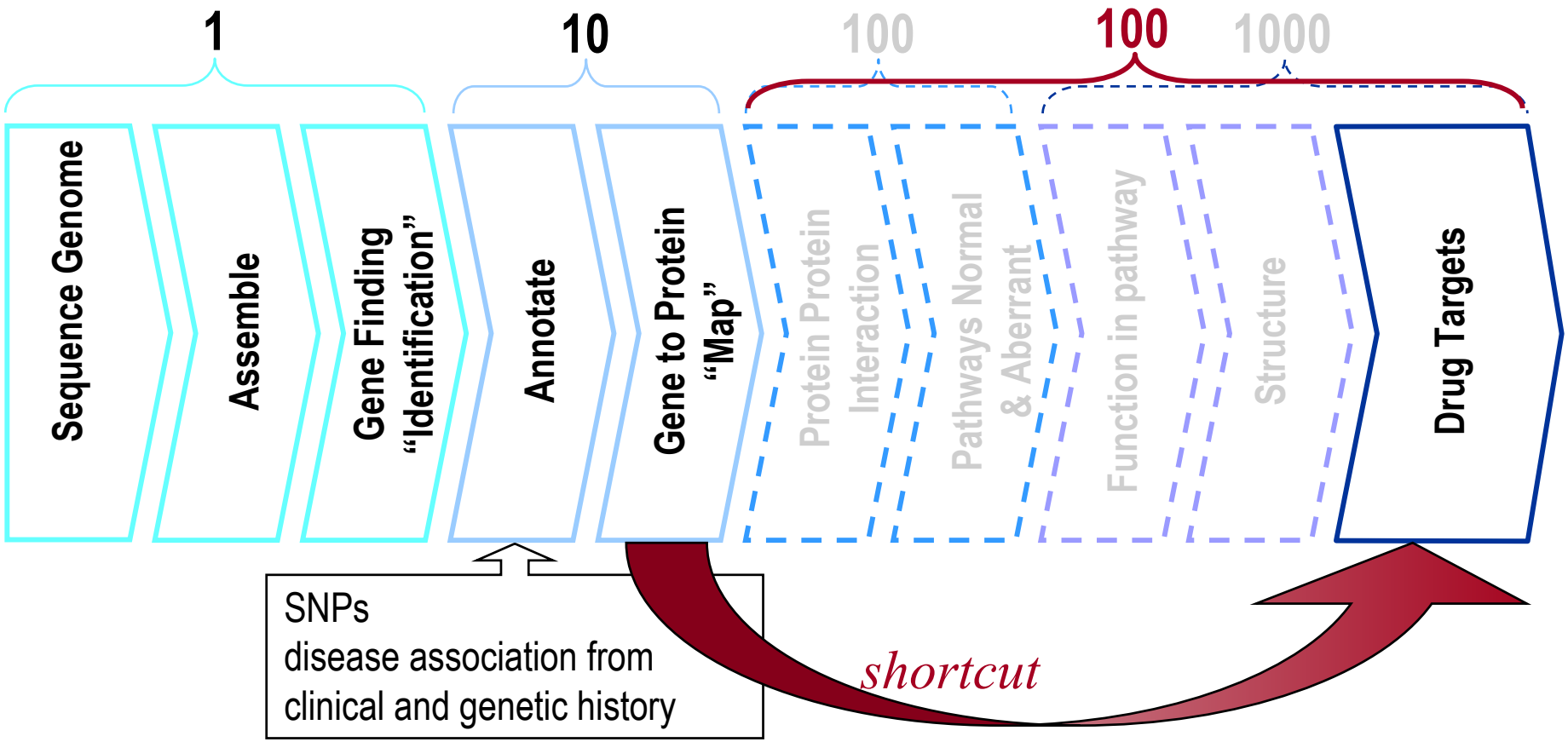
Today

2020--2030?

Bill Camp
Bill@sandia.gov

Sandia National Laboratories

…High-Throughput Experiments can create a *shortcut*

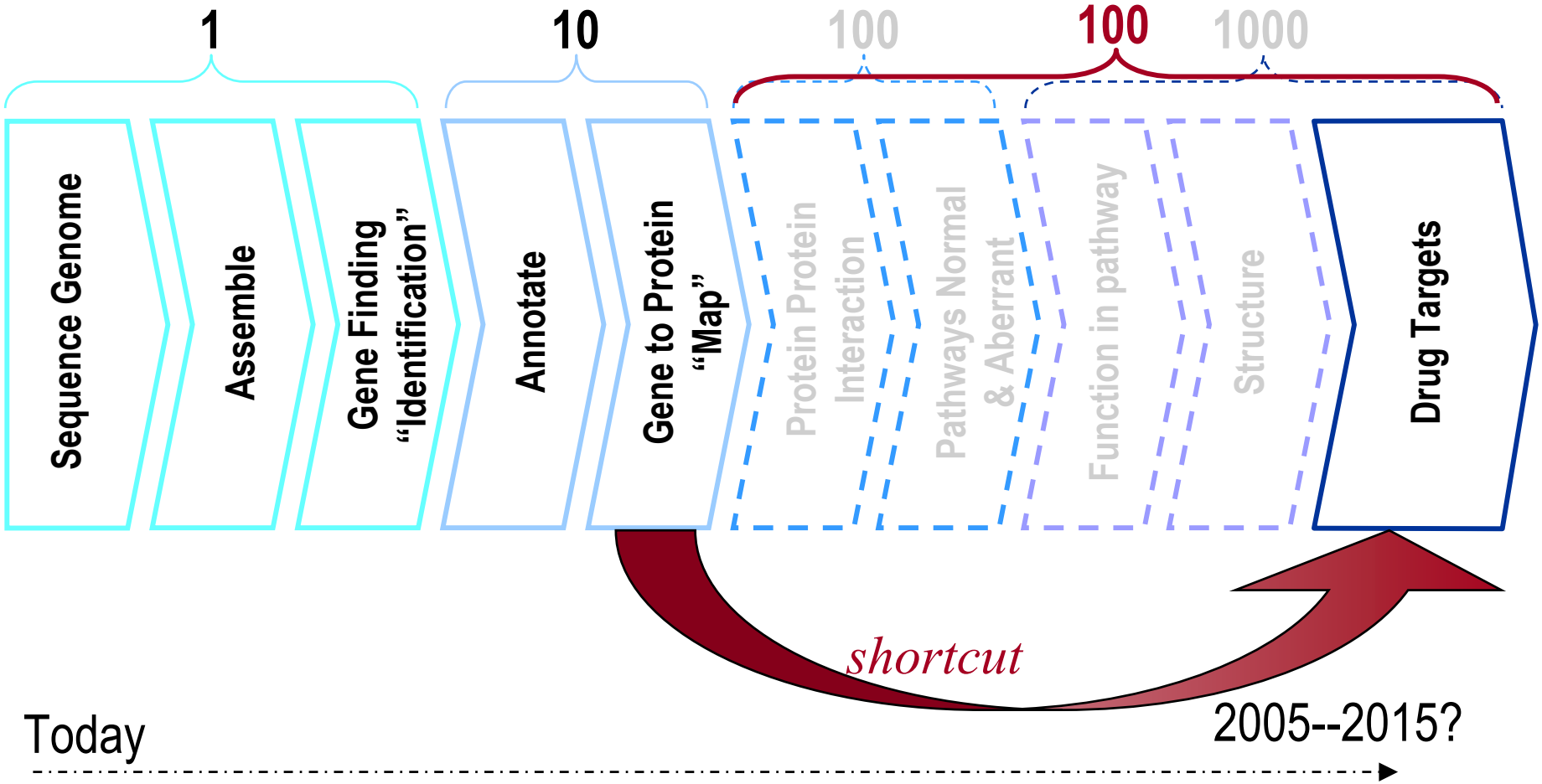


Existing clinical data and tissue banks could be CRITICAL because …

Bill Camp
Bill@sandia.gov

Sandia National Laboratories

Computing Power Needs May Decrease with Clinical Collaborations



Bill Camp
Bill@sandia.gov

Sandia National Laboratories

Big Pharma (& biotech) are increasingly driving the high-end computing market.

Annual Sales	$300B
Historic Growth Rate	10-14%
R&D Expenditures	$62B ($26.4B in US)
	$4.6B external R&D to understand human genome
	11% of sales in 1980
	20% of sales in 2000

- 70% of patented drugs come off patent in the next 4 years.
- 80% average drop in sales revenue when patent expires.
- $600M average drug development cost.
- Diminishing pool of easy targets.

Bill Camp
Bill@sandia.gov

Sandia National Laboratories

The Implications of the New Biology for High-End Computing are Growing

IT market for life sciences forecast to reach $40B by 2004, e.g.

Celera builds 1st tera-cluster for biotechnology– speeds up genomics by 10x

IBM, Compaq: $100 million investments in the life sciences market

NuTec 7.5 Tflops IBM cluster (US, & Europe-planned)

GeneProt Large-Scale Proteomic Discovery And Production Facility:1,420 Alpha processors.

Blackstone Linux/Intel Clusters (Pfizer, Biogen, AstraZeneca, & 10-15 more on the way)

And so on ...

Sandia's strategy: Partner with the best to lead development of scientific supercomputing solutions.

Bill Camp
Bill@sandia.gov

Sandia National Laboratories